Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10972–10976, December 1993
Genetics

Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci

(gene mapping/molecular genetic markers/quantitative genetics/multiple regression/interval test)

ZHAO-BANG ZENG

Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

Communicated by C. Clark Cockerham, August 18, 1993 (received for review June 11, 1993)

ABSTRACT **It is now possible to use complete genetic linkage maps to locate major quantitative trait loci (QTLs) on chromosome regions. The current methods of QTL mapping (e.g., interval mapping, which uses a pair or two pairs of flanking markers at a time for mapping) can be subject to the effects of other linked QTLs on a chromosome because the genetic background is not controlled. As a result, mapping of QTLs can be biased, and the resolution of mapping is not very high. Ideally when we test a marker interval for a QTL, we would like our test statistic to be independent of the effects of possible QTLs at other regions of the chromosome so that the effects of QTLs can be separated. This test statistic can be constructed by using a pair of markers to locate the testing position and at the same time using other markers to control the genetic background through a multiple regression analysis. Theory is developed in this paper to explore the idea of a conditional test via multiple regression analysis. Various properties of multiple regression analysis in relation to QTL mapping are examined. Theoretical analysis indicates that it is advantageous to construct such a testing procedure for mapping QTLs and that such a test can potentially increase the precision of QTL mapping substantially.**

Lander and Botstein (1) proposed an interval method to map for major quantitative trait loci (QTLs) systematically in a genome in experimental organisms. There are several advantages of their method compared with traditional regression analysis (2). There are, however, still some problems with Lander–Botstein's interval mapping method, as follows. (*i*) The test statistic on an interval can be affected by QTLs located at other regions of the chromosome. Even though there is no QTL within an interval, the test statistic on the interval can still be very significant if there is a QTL at some nearby point on the chromosome. Moreover, if there is more than one QTL on a chromosome, the test statistic at a testing position will be affected by all those QTLs, and the estimated positions and effects of "QTLs" identified by this method are likely to be biased (3, 4). (*ii*) It is not efficient to use only two markers at a time to do the test, as the information from other markers is not utilized. Lander and Botstein (1) also proposed a simultaneous search strategy for multiple QTLs on multiple intervals to alleviate some of these problems. But as the search becomes multidimensional (1, 3), there are some difficulties in parameter estimation and model identifiability. Both effort and ambiguity can be multiplied. Besides, the number of QTLs on a chromosome is unknown, and mapping can still be biased. Also, information in the data from other markers is still not utilized.

Ideally, when we test an interval for a QTL, we would like to have our test statistic independent of the effects of possible QTLs located at other regions of the chromosome. By so doing, we can eliminate biases in our mapping from those QTLs and potentially increase the precision of mapping. Such a test statistic can be constructed by combining Lander–Botstein's interval mapping with multiple regression analysis. In this paper, theoretical implications of multiple regression analysis in relation to QTL mapping are explored. It is shown that the partial regression coefficient of the phenotype on a marker in multiple regression depends only on those QTLs that are located in the interval bracketed by the two neighboring markers and is independent of QTLs located in other intervals. Then, using this property, we can construct a test statistic that is independent of effects of QTLs in other regions of the chromosomes. This result provides a basis for constructing an interval test for mapping QTLs. Also, by fitting multiple markers in a regression model, much background genetic variation in a population can be controlled in analysis, and, as a result, statistical power of detecting QTLs can be improved. The advantages and disadvantages of fitting multiple markers in the model for mapping QTLs are discussed, as are procedures to construct an appropriate interval test for mapping QTLs.

PROPERTIES OF MULTIPLE REGRESSION ANALYSIS

The Model. Let us consider, for simplicity, a backcross population that is from two inbred parental populations, P_1 and P_2, fixed for different alleles at m QTLs and t markers. Let the means of the P_1 and P_2 populations be μ_1 and μ_2 and the difference between μ_1 and μ_2 be $\mu_1 - \mu_2 = \sum_{u=1}^{m} c_u$, ignoring epistasis, where c_u is the effect difference between the two homozygotes for alleles fixed in the two parental populations at the uth QTL. The value of c_u can be positive or negative. The mean of the F_1 population, which is a cross between P_1 and P_2, is then defined as $\mu_{F_1} = \mu_2 + \sum_{u=1}^{m} \frac{1}{2}(1 + d_u)c_u$, where d_u is the degree of dominance at the uth QTL. The trait values of individuals in a backcross population between P_1 and F_1 are defined as $y_h = \mu_2 + \sum_{u=1}^{m} [\xi_{uh}c_u + \frac{1}{2}(1 - \xi_{uh})(1 + d_u)c_u] + e_h$, where y_h is the trait value of the hth individual, ξ_{uh} is an indicator variable taking a value 1 or 0 with equal probability, and e_h is a random environmental deviation with mean zero and variance σ_e^2. If the uth and vth QTLs are unlinked, ξ_{uh} and ξ_{vh} are independent; otherwise ξ_{uh} and ξ_{vh} are correlated with the correlation coefficient $(1 - 2r_{uv})$, where r_{uv} is the recombination frequency between the uth and vth QTLs. The variance of the trait value y in the backcross population is then $\sigma_y^2 = \sigma_e^2 + \frac{1}{4}\sum_{u=1}^{m} a_u^2 + \frac{1}{4}\sum_{v=1}^{m} \sum_{u=1,u\neq v}^{m} (1 - 2r_{uv})a_u a_v$, where $a_u = (1 - d_u)c_u/2$ is the effect of the uth QTL expressed as a difference in effects between the homozygote in P_1 and the heterozygote in F_1. The double summation represents the covariance among loci due to linkage disequilibrium.

Abbreviation: QTLs, quantitative trait loci.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Suppose that we have a sample of n individuals from this backcross population and observations on the quantitative trait and t ordered markers. One way to analyze these data is to perform multiple regression with a linear model $y_h = b_0 + \sum_{i=1}^{t} b_i x_{hi} + \varepsilon_h$ for $h = 1, 2, \ldots, n$, where x_{hi} is the type of the ith marker in the hth individual, taking a value 1 or 0 depending on whether the marker is homozygous or heterozygous; b_0 is the mean of the model; b_i (also denoted by $b_{yi \cdot s_i}$, where s_i denotes a set that includes all markers except the ith marker) is the partial regression coefficient of the phenotype y on the ith marker conditional on all other markers; and ε_h is a random variable. In this paper, the subscripts u, v, and q are used for indexing QTLs; i, j, k, and l are used for markers; and h is used for individuals.

As will become clear, however, this analysis is not appropriate for mapping QTLs in itself. But there are several distinctive features of multiple regression analysis that can be used to devise a mapping method to improve the precision and efficiency of QTL mapping. In the following sections, I explore the advantages and disadvantages of multiple regression analysis by analyzing the expected values of the regression coefficients, the expected sampling variances of the coefficients, and the sampling correlation of the coefficients in terms of genetic parameters.

Partial Regression Coefficient. To analyze the partial regression coefficient $b_{yi \cdot s_i} (= \sigma_{yi \cdot s_i}/\sigma^2_{i \cdot s_i})$, we need to analyze the conditional variance $\sigma^2_{i \cdot s_i}$ and the conditional covariance $\sigma_{yi \cdot s_i}$. Since it is a backcross population, the variance of each marker variable in the population is $\sigma_i^2 = 1/4$ for $i = 1, 2, \ldots, t$. It is easy to show that the covariance between the ith and jth markers is $\sigma_{ij} = (1 - 2r_{ij})/4$, where r_{ij} is the recombination frequency between the two markers. The covariance between the trait value y and the ith marker is $\sigma_{yi} = \sum_{u=1}^{m}(1 - 2r_{ui})a_u/4$, where r_{ui} is the recombination frequency between the uth QTL and ith marker.

With these basic equations, any conditional variance and covariance can be derived (5). First note that under Haldane's mapping function, assuming no interference, the recombination frequency between markers i and j can be expressed as $(1 - 2r_{ij}) = (1 - 2r_{ik})(1 - 2r_{kj})$ for $i < k < j$. Here, for simplicity, it is assumed that all markers in a genome are numbered according to physical order from the first chromosome to the last chromosome, so that the relation or ordering $i < k < j$ indicates that marker k is located between markers i and j. For two markers on different chromosomes, the relative order is immaterial.

The variance of marker i conditional on marker j is then $\sigma^2_{i \cdot j} = \sigma_i^2 - \sigma_{ij}^2/\sigma_j^2 = [1 - (1 - 2r_{ij})^2]/4 = r_{ij}(1 - r_{ij})$. The covariance between markers i and j conditional on marker k is $\sigma_{ij \cdot k} = \sigma_{ij} - \sigma_{ik}\sigma_{jk}/\sigma_k^2 = [(1 - 2r_{ij}) - (1 - 2r_{ik})(1 - 2r_{jk})]/4$, which is 0 if $i < k < j$ or $j < k < i$, is $r_{jk}(1 - r_{jk})(1 - 2r_{ij})$ if $i < j < k$ or $k < j < i$, and is $r_{ik}(1 - r_{ik})(1 - 2r_{ij})$ if $k < i < j$ or $j < i < k$. By taking $i < j$ and $k < l$, the covariance between markers i and j conditional on markers k and l is $\sigma_{ij \cdot kl} = \sigma_{ij \cdot k} - \sigma_{il \cdot k}\sigma_{jl \cdot k}/\sigma^2_{l \cdot k} = \sigma_{ij \cdot l} - \sigma_{ik \cdot l}\sigma_{jk \cdot l}/\sigma^2_{k \cdot l}$, which is 0 if $i < k < j < l$ or $i < k < l < j$ or $k < i < l < j$, is $\sigma_{ij \cdot k}$ if $i < j < k < l$, is $\sigma_{ij \cdot l}$ if $k < l < i < j$, and is $[r_{ki}(1 - r_{ki})r_{jl}(1 - r_{jl})(1 - 2r_{ij})]/[r_{kl}(1 - r_{kl})]$ if $k < i < j < l$. From this analysis, it is immediately clear that if $k < i < j < l$, the covariance between markers i and j conditional on markers $k - 1$, k, l, and $l + 1$ is $\sigma_{ij,(k-1)kl(l+1)} = \sigma_{ij,kl(l+1)} = \sigma_{ij,kl}$. In general, the covariance between markers i and $i + 1$ conditional on all other markers is $\sigma_{i(i+1) \cdot s_{i(i+1)}} = \sigma_{i(i+1) \cdot (i-1)(i+2)}$, where $s_{i(i+1)}$ denotes a set that includes all markers except markers i and $i + 1$.

It is very important to note that, conditional on an intermediate marker, the covariance between the two flanking markers is zero, as shown, for example, by $\sigma_{ij \cdot k} = 0$ if $i < k < j$. Also note that conditioning on a flanking marker, say k, makes the covariance between markers i and j independent of all those markers (and QTLs, see below) that are located at the other side of marker k, as shown, for example, by $\sigma_{ij \cdot kl} = \sigma_{ij \cdot k}$ if $i < j < k < l$. Thus, conditioning on two flanking markers, say $i - 1$ and $i + 2$, makes the covariance between two interior markers i and $i + 1$ independent of all those markers (and QTLs, see below) that are outside the marker interval $(i - 1, i + 2)$. These are properties of the linear ordering of markers and QTLs on chromosomes and the very basis for the interval test.

Also, by taking $j < k$, the variance of marker i conditional on markers j and k is $\sigma^2_{i \cdot jk} = \sigma^2_{i \cdot j} - \sigma^2_{ik \cdot j}/\sigma^2_{k \cdot j} = \sigma^2_{i \cdot k} - \sigma^2_{ij \cdot k}/\sigma^2_{j \cdot k}$, which is $\sigma^2_{i \cdot j}$ if $i < j < k$, is $\sigma^2_{i \cdot k}$ if $j < k < i$, and is $[r_{ji}(1 - r_{ji})r_{ik}(1 - r_{ik})]/[r_{jk}(1 - r_{jk})]$ if $j < i < k$. It is not difficult to see that, in general, $\sigma^2_{i \cdot s_i} = \sigma^2_{i \cdot (i-1)(i+1)}$.

The covariance between y and marker i conditional on marker j is a little more complicated:

$$\sigma_{yi \cdot j} = \sigma_{yi} - \sigma_{yj}\sigma_{ij}/\sigma_j^2$$

$$= \frac{1}{4}\sum_{u=1}^{m}[(1 - 2r_{ui}) - (1 - 2r_{uj})(1 - 2r_{ij})]a_u$$

$$= \begin{cases} r_{ij}(1 - r_{ij})\sum_{u \le i}(1 - 2r_{ui})a_u \\ \quad + \sum_{i<u<j} r_{uj}(1 - r_{uj})(1 - 2r_{iu})a_u & \text{if } i < j \\ r_{ij}(1 - r_{ij})\sum_{u \ge i}(1 - 2r_{iu})a_u \\ \quad + \sum_{j<u<i} r_{ju}(1 - r_{ju})(1 - 2r_{ui})a_u & \text{if } j < i. \end{cases}$$

It is clear that $\sigma_{yi \cdot jk} = \sigma_{yi \cdot j} - \sigma_{yk \cdot j}\sigma_{ik \cdot j}/\sigma^2_{k \cdot j} = \sigma_{yi \cdot k} - \sigma_{yj \cdot k}\sigma_{ij \cdot k}/\sigma^2_{j \cdot k}$, which is $\sigma_{yi \cdot j}$ if $i < j < k$ and $\sigma_{yi \cdot k}$ if $i < k < j$. If $k < i < j$, it can be shown after some analysis that

$$\sigma_{yi \cdot jk} = \frac{r_{ij}(1 - r_{ij})}{r_{kj}(1 - r_{kj})}\sum_{k<u\le i} r_{ku}(1 - r_{ku})(1 - 2r_{ui})a_u + \frac{r_{ki}(1 - r_{ki})}{r_{kj}(1 - r_{kj})}\sum_{i<u<j} r_{uj}(1 - r_{uj})(1 - 2r_{iu})a_u.$$

Again, in general, $\sigma_{yi \cdot s_i} = \sigma_{yi \cdot (i-1)(i+1)}$.

Therefore, the partial regression coefficient is

$$b_{yi \cdot s_i} = \frac{\sigma_{yi \cdot s_i}}{\sigma^2_{i \cdot s_i}} = \frac{\sigma_{yi \cdot (i-1)(i+1)}}{\sigma^2_{i \cdot (i-1)(i+1)}}$$

$$= \sum_{i-1<u\le i} \frac{r_{(i-1)u}(1 - r_{(i-1)u})(1 - 2r_{ui})}{r_{(i-1)i}(1 - r_{(i-1)i})}a_u + \sum_{i<u<i+1} \frac{r_{u(i+1)}(1 - r_{u(i+1)})(1 - 2r_{iu})}{r_{i(i+1)}(1 - r_{i(i+1)})}a_u, \qquad [1]$$

where the first summation is for all QTLs located between markers $i - 1$ and i, and the second summation is for all QTLs located between markers i and $i + 1$. This regression coefficient depends only on those QTLs which are located between markers $i - 1$ and $i + 1$. This is a very desirable property. By using this property we can create an interval test in which we can test whether there is a QTL within a marker interval. Qualitatively this conditional test is more precise than unconditional tests in which we can ask only whether there is a QTL on a chromosome.

If there is only one QTL, u, between markers $i - 1$ and $i + 1$ (say between markers $i - 1$ and i), then $b_{yi \cdot s_i} = \{[r_{(i-1)u}(1 - r_{(i-1)u})(1 - 2r_{ui})]/[r_{(i-1)i}(1 - r_{(i-1)i})]\}a_u \approx [r_{(i-1)u}/r_{(i-1)i}]a_u$, for small $r_{(i-1)i}$ (say, <0.2). An estimate $\hat{b}_{yi \cdot s_i}$ of $b_{yi \cdot s_i}$ will then be a biased estimate of a_u unless QTL u is located right on the marker i. The ratio $r_{(i-1)u}/r_{(i-1)i}$ changes from 0 to 1 monotonically and almost linearly as QTL u shifts its location from

marker $i-1$ to marker i. Statistically, what this conditional test does is to make the marker interval $(i-1, i+1)$ act essentially like a chromosome.

Sampling Variance of the Partial Regression Coefficient. To analyze the sampling variance of the regression coefficient, we need the conditional phenotypic variance. This variance is (5):

$$\begin{aligned}
\sigma^2_{y\cdot s} &= \sigma^2_y - \sum_{i=1}^{t} b_{yi\cdot s_i}\sigma_{yi} = \sigma^2_y - \sum_{i=1}^{t}\left[\left(\sum_{i-1<v<i}\frac{r_{(i-1)v}(1-r_{(i-1)v})(1-2r_{vi})}{r_{(i-1)i}(1-r_{(i-1)i})}a_v\right.\right.\\
&\qquad\left.\left.+\sum_{i<v<i+1}\frac{r_{v(i+1)}(1-r_{v(i+1)})(1-2r_{iv})}{r_{i(i+1)}(1-r_{i(i+1)})}a_v\right)\frac{1}{4}\sum_{u=1}^{m}(1-2r_{ui})a_u\right]\\
&= \sigma^2_e + \frac{1}{4}\sum_{u=1}^{m}a^2_u + \frac{1}{4}\sum_{v=1}^{m}\sum_{u=1,u\neq v}^{m}(1-2r_{uv})a_ua_v - \frac{1}{4}\sum_{i=1}^{t}\left[\sum_{i-1<v\leq i}\frac{r_{(i-1)v}(1-r_{(i-1)v})(1-2r_{vi})^2}{r_{(i-1)i}(1-r_{(i-1)i})}a^2_v\right.\\
&\qquad+\sum_{i<v<i+1}\frac{r_{v(i+1)}(1-r_{v(i+1)})(1-2r_{iv})^2}{r_{i(i+1)}(1-r_{i(i+1)})}a^2_v\\
&\qquad+\sum_{i-1<v<i}\sum_{u=1,u\neq v}^{m}\frac{r_{(i-1)v}(1-r_{(i-1)v})(1-2r_{vi})(1-2r_{ui})}{r_{(i-1)i}(1-r_{(i-1)i})}a_ua_v\\
&\qquad\left.+\sum_{i<v<i+1}\sum_{u=1,u\neq v}^{m}\frac{r_{v(i+1)}(1-r_{v(i+1)})(1-2r_{iv})(1-2r_{ui})}{r_{i(i+1)}(1-r_{i(i+1)})}a_ua_v\right]\\
&= \sigma^2_e + \frac{1}{4}\sum_{u=1}^{m}\left[1-\frac{r_{u_Lu}(1-r_{u_Lu})(1-2r_{uu_R})^2 + r_{uu_R}(1-r_{uu_R})(1-2r_{u_Lu})^2}{r_{u_Lu_R}(1-r_{u_Lu_R})}\right]a^2_u\\
&\quad+\frac{1}{4}\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left[2(1-2r_{uv})-\frac{r_{u_Lu}(1-r_{u_Lu})(1-2r_{uu_R})(1-2r_{u_Rv})}{r_{u_Lu_R}(1-r_{u_Lu_R})}\right.\\
&\qquad-\frac{r_{uu_R}(1-r_{uu_R})(1-2r_{u_Lu})(1-2r_{u_Lv})}{r_{u_Lu_R}(1-r_{u_Lu_R})}-\frac{r_{v_Lv}(1-r_{v_Lv})(1-2r_{vv_R})(1-2r_{uv_R})}{r_{v_Lv_R}(1-r_{v_Lv_R})}\\
&\qquad\left.-\frac{r_{vv_R}(1-r_{vv_R})(1-2r_{v_Lv})(1-2r_{uv_L})}{r_{v_Lv_R}(1-r_{v_Lv_R})}\right]a_ua_v, \qquad [2]
\end{aligned}$$

where s is the set of all markers. In the above equation, the linkage relations between markers and QTLs are indicated by summations like $\Sigma_{i<v<i+1}$, which denotes summation for all QTLs that are located between markers i and $i+1$ and by notations like u_L and u_R, which refer to the markers located on the immediate left and right of the uth QTL. The first summation term in Eq. 2 contains the residual genetic variance within QTLs after fitting all available markers, and the second summation term contains the residual genetic covariance between QTLs. If there is at most one QTL for each marker interval (which means that there exists at least one marker between any two QTLs so that for all $u < v$, the right flanking marker of QTL u is always on the left side of QTL v (i.e., $u_R < v$ and $u < v_L$), the second summation term is zero, and there will be no residual genetic covariance among QTLs in the conditional variance because in this case all QTLs are physically separated by markers and made to be independent from each other by conditioning on those markers. Then $\sigma^2_{y\cdot s} = \sigma^2_e + \Sigma_{u=1}^{m}\{[r_{u_Lu}(1-r_{u_Lu})r_{uu_R}(1-r_{uu_R})]/[r_{u_Lu_R}(1-r_{u_Lu_R})]\}a^2_u$.

If, for example, markers are evenly spaced throughout a genome with a recombination frequency r for each marker interval (not for those implied intervals linking two chromosomes), the conditional variance will be at most $\sigma^2_{y\cdot s} = \sigma^2_e + \Sigma_{u=1}^{m}\{r/[4(1-r)]\}a^2_u$ (assuming that all QTLs are located in the middle of marker intervals, which is the worst situation). Thus, the genetic part of the variance $\sigma^2_{y\cdot s}$ is reduced by at least a factor of $(1-2r)/r$ compared with that of the variance σ^2_y. Taking linkage disequilibrium in σ^2_y into account will further increase this reduction.

Now the sampling variance of the regression coefficient for a large sample is (5) approximately

$$\begin{aligned}
\mathrm{Var}(\hat{b}_{yi\cdot s_i}) &\simeq \frac{\sigma^2_{y\cdot s}}{n\sigma^2_{i\cdot s_i}}\\
&= \frac{1}{n}\left[\sigma^2_e + \sum_{u=1}^{m}\frac{r_{u_Lu}(1-r_{u_Lu})r_{uu_R}(1-r_{uu_R})}{r_{u_Lu_R}(1-r_{u_Lu_R})}a^2_u\right]\\
&\quad\div\frac{r_{(i-1)i}(1-r_{(i-1)i})r_{i(i+1)}(1-r_{i(i+1)})}{r_{(i-1)(i+1)}(1-r_{(i-1)(i+1)})}. \qquad [3]
\end{aligned}$$

It is of interest to compare Eq. 3 to the sampling variance of the simple regression coefficient in the linear model $y_h = b_0 + b_ix_{hi} + \varepsilon_h$ for $h = 1, 2, \ldots, n$. By assuming that there is only one QTL, q, which is linked to marker i, the sampling

variance of the simple regression coefficient of trait value y on marker i (5) is

$$\mathrm{Var}(\hat{b}_{yi}) \simeq \frac{\sigma^2_{y\cdot i}}{n\sigma^2_i} = \frac{4}{n}\left[\sigma^2_e + r_{iq}(1-r_{iq})a^2_q + \frac{1}{4}\sum_{u=1,u\neq q}^{m} a^2_u + \frac{1}{4}\sum_{u=1}^{m}\sum_{v=1,v\neq u}^{m}(1-2r_{uv})a_u a_v\right].$$

Since $\sigma^2_{y\cdot s} < \sigma^2_{y\cdot i}$ and $\sigma^2_{i\cdot s_i} \leq \sigma^2_i$, $\mathrm{Var}(\hat{b}_{yi\cdot s_i})$ may be smaller or larger than $\mathrm{Var}(\hat{b}_{yi})$, depending on the orders of reduction of $\sigma^2_{y\cdot s}$ compared to $\sigma^2_{y\cdot i}$ and $\sigma^2_{i\cdot s_i}$ compared to σ^2_i. For multiple QTLs the genetic part of the variance $\sigma^2_{y\cdot s}$ is very roughly about r times the genetic part of the variance $\sigma^2_{y\cdot i}$, where r is the average recombination frequency of marker intervals. Depending on heritability, this reduction can be very substantial. On the other hand, $\sigma^2_{i\cdot s_i}$ will be smaller than σ^2_i if $r_{(i-1)i} < 0.5$ or $r_{i(i+1)} < 0.5$, and, as a result, $\mathrm{Var}(\hat{b}_{yi\cdot s_i})$ may be larger than $\mathrm{Var}(\hat{b}_{yi})$.

However, since $\sigma^2_{i\cdot s_i}$ is a function of only the sizes of the neighboring intervals [i.e., the intervals $(i-1, i)$ and $(i, i+1)$ in Eq. 3], deliberately removing some immediately linked markers in the analysis [such as removing markers $i-1$ and $i+1$ so that $\sigma^2_{i\cdot s_{(i-1)(i+1)}}$ is a function of $r_{(i-2)i}$, $r_{i(i+2)}$ and $r_{(i-2)(i+2)}$] will increase the denominator of the sampling variance (Eq. 3) and at the same time minimize the numerator of the variance, thus minimizing the sampling variance of the partial regression coefficient and increasing the statistical power of testing. This is because the conditional test on $\hat{b}_{yi\cdot s_i}$ relies on the number of recombinants with recombination that occurred between markers $i-1$ and i or between markers i and $i+1$. When the intervals $(i-1, i)$ and $(i, i+1)$ are small, there will be few such recombinants, and the effective sample size for the test will be small. Thus, increasing the sizes of the neighboring intervals in the test (i.e., testing $\hat{b}_{yi\cdot s_{(i-1)(i+1)}}$ for example) will increase the number of recombinants and thus increase the statistical power of the test. But removing some immediately linked markers, such as markers $i-1$ and $i+1$, in the model will increase the chance of interference of possible linked multiple QTLs on the interval $(i-2, i+2)$ on hypothesis testing and parameter estimation.

Sampling Correlation of the Partial Regression Coefficient. It can be shown that the sampling correlation of the partial regression coefficients, $\hat{b}_{yi\cdot s_i}$ and $\hat{b}_{yj\cdot s_j}$ $(i < j$ and assuming $r_{ij} > 0)$, is expected to be

$$\mathrm{Corr}(\hat{b}_{yi\cdot s_i}, \hat{b}_{yj\cdot s_j}) = -\gamma_{ij\cdot s_{ij}} = -\frac{\sigma_{ij\cdot s_{ij}}}{\sigma_{i\cdot s_{ij}}\sigma_{j\cdot s_{ij}}} = \begin{cases} -(1-2r_{i(i+1)})\left[\dfrac{r_{(i-1)i}(1-r_{(i-1)i})r_{(i+1)(i+2)}(1-r_{(i+1)(i+2)})}{r_{i(i+2)}(1-r_{i(i+2)})r_{(i-1)(i+1)}(1-r_{(i-1)(i+1)})}\right]^{1/2} & \text{if } j = i+1 \\ 0 & \text{otherwise,} \end{cases} \quad [4]$$

where $\gamma_{ij\cdot s_{ij}}$ is the partial correlation coefficient between markers i and j conditional on all other markers. This shows that the sampling correlation of the partial regression coefficients of the phenotype on markers i and j is generally expected to be zero unless the two markers are adjacent markers.

This correlation is closely related to the correlation of the usual F test statistics for $\hat{b}_{yi\cdot s_i}$ and $\hat{b}_{yi\cdot s_j}$ individually. For example, if we want to test the hypotheses $H_0: b_{yi\cdot s_i} = 0$ and $H_1: b_{yi\cdot s_i} \neq 0$, we can use the F statistic: $F_i = \hat{b}^2_{yi\cdot s_i}/\widehat{\mathrm{Var}}(\hat{b}_{yi\cdot s_i})$, where $\widehat{\mathrm{Var}}(\hat{b}_{yi\cdot s_i})$ is an estimate of the sampling variance of $\hat{b}_{yi\cdot s_i}$. Statistic F_i has an F distribution with 1 and $n-t-1$ degrees of freedom. When the sample size n is large (significantly larger than the number of markers, t, fitted in the model), the distribution of F_i is approximately χ^2 with 1 degree of freedom. If ε_h values in the regression model are normally distributed, $\hat{b}_{yi\cdot s_i}$ values are also (multivariately) normally distributed with means, variances, and correlations given by Eqs. **1**, **3**, and **4**. Then, from properties of the multivariate normal distribution, the correlation between F_i and F_j is $\mathrm{Corr}(F_i, F_j) \rightarrow \mathrm{Corr}(\hat{b}^2_{yi\cdot s_i}, \hat{b}^2_{yj\cdot s_j}) = \mathrm{Corr}(\hat{b}_{yi\cdot s_i}, \hat{b}_{yj\cdot s_j})^2/2 = \gamma^2_{ij\cdot s_{ij}}/2$ as $n \rightarrow \infty$, which is zero unless markers i and j are adjacent markers.

DISCUSSION

In summary, in relation to QTL mapping, multiple regression analysis has the following properties. (*i*) If there is no epistasis, the partial regression coefficient of the trait on a marker depends only on those QTLs that are located in the interval bracketed by the two neighboring markers and is independent of QTLs located in other intervals. This is the basis for an interval test. (*ii*) Conditioning on unlinked markers in the analysis will reduce the sampling variance of the test statistic by controlling some residual genetic variation and thus will increase statistical power of the test. This useful information has not been utilized by the current QTL mapping methods (1). (*iii*) Conditioning on linked markers in the analysis will reduce the chance of interference of possible multiple linked QTLs on hypothesis testing and parameter estimation and thus potentially increase the precision of the test and estimation, but with a possible decrease of statistical power of the test. This summarizes the advantages and disadvantages of the interval test: that is, there is a trade-off between precision and efficiency of the mapping by using an interval test. (*iv*) Two sample partial regression coefficients of the trait y on markers i and j, $\hat{b}_{yi\cdot s_i}$ and $\hat{b}_{yj\cdot s_j}$, are generally uncorrelated unless the two markers i and j are adjacent markers. This is related to the correlation of test statistics between two testing positions in two intervals for an interval test and is indirectly related to the issue of determining an appropriate significance value of a test statistic under a null hypothesis for an overall test covering a whole genome.

So far the theoretical analysis has been presented for a backcross population. It can be shown that similar conclusions and properties of multiple regression analysis in relation to QTL mapping hold for other population designs, such as the F_2 population intercrossed from F_1. For example, if there is no dominance at the m QTLs (i.e., all $d_u = 0$), the phenotypic variance of the F_2 population will be that of the backcross population with the genetic variance and covariance multiplied by a factor of two. In F_2, there are three possible genotypes for each marker. If we let x_{hi} in the

statistical model take values 2, 1, or 0 if the genotype of the ith marker in the hth individual is P_1 homozygote, heterozygote, or P_2 homozygote, then σ_i^2, σ_{ij}, and σ_{yi} will be just twice those for the backcross population. Then the results of Eqs. 1 and 4 are unchanged. Both the denominator $\sigma^2_{i \cdot s_i}$ and the genetic part of the numerator $\sigma^2_{y \cdot s}$ of Eq. 3 are also multiplied by 2, so that the sampling variance of $\hat{b}_{yi \cdot s_i}$ in Eq. 3 will be decreased. Basically the above four properties are unchanged. Qualitatively, this also applies if we take dominance into account, although the analysis with dominance in F_2 will be more complicated and the regression model may need to introduce one more variable for each marker to accommodate dominance deviation. With epistasis, however, property 1 will not be true exactly. Depending on the type and degree of epistasis, property 1 may hold approximately in some cases.

The four properties provide the theoretical basis for constructing an interval test. Direct use of multiple regression analysis for QTLs mapping is, however, inappropriate and inefficient as the partial regression coefficient is more than likely to be a biased estimate of the relevant QTL effect. It would be appropriate to combine Lander–Botstein's interval mapping with multiple regression to construct an interval test for testing and estimating QTL effects. Detailed testing procedures for such an interval test depend on genetic models, experimental designs, and data structures. For example, for a backcross population we can use the following statistical model to test for a QTL on a marker interval $(i, i + 1)$: $y_h = b_0 + b^* x^*_h + \Sigma_{j \neq i, i+1} b_j x_{hj} + \varepsilon_h$ for $h = 1, 2, \ldots, n$, where b^* is the effect of the putative QTL expressed as a difference in effects between the homozygote and heterozygote and x^*_h is an indicator variable, taking a value 1 or 0 with probability depending on the genotypes of the markers i and $i + 1$ and the testing position of the putative QTL. Statistically, this is a mixture model with the mixing proportions (i.e., the probability of $x^*_j = 1$) 1, $p(= r_{iq}/r_{i(i+1)})$, $1 - p$, and 0 for the four different genotypes of the markers i and $i + 1$ (homozygote/homozygote, heterozygote/homozygote, homozygote/heterozygote, and heterozygote/heterozygote, respectively), ignoring double recombination between markers i and $i + 1$, where r_{iq} is the recombination frequency between marker i and the putative QTL, q, and $r_{i(i+1)}$ is the recombination frequency between markers i and $i + 1$. With an appropriate assumption about the distribution of the random variable ε, a maximum likelihood ratio test statistic can be constructed and computed for the hypotheses $H_0 : b^* = 0$ and $H_1 : b^* \neq 0$ (the detailed procedure will be discussed elsewhere). As explained above, this test is an interval test with the test statistic unaffected by QTLs located outside the marker interval $(i - 1, i + 2)$ if markers $i - 1$ and $i + 2$ are fitted in the model along with other markers. This test can be performed at any position in a genome covered by markers just as for the interval mapping of Lander–Botstein. But when comparing this test with the interval mapping of Lander–Botstein, this test is more likely to detect and estimate the effect of a single QTL at any testing position because it is an interval test. This would then create a convenient systematic searching strategy for multiple QTLs as it reduces a multidimensional search problem (1, 3) (for multiple QTLs on a chromosome) to a one-dimensional search problem. Also, by the virtue of maximum likelihood principles, estimates of QTL positions and effects by this method will tend to be asymptotically unbiased.

There are, however, two issues that need to be addressed for this mapping method. First, since it is a multiple test and search problem (for multiple locations), what would be an appropriate significance value for the test statistic given a data set? Lander and Botstein (1) discussed the issue of the appropriate significance value of the test statistic (using the logarithm of odds score) for their mapping procedure covering a whole genome. The threshold of the test statistic for testing the null hypothesis for this method is, however, different. The difference is that with multiple regression the test statistics are almost independent between different intervals, as indicated above, but highly correlated within intervals. The implications of these properties and practical determination of an appropriate significance value of the test statistic given a data set will be discussed elsewhere. Second, what would be an optimum model for QTL mapping? Or, how many and what markers should be included in the model as a background control? Properties 2 and 3 briefly summarize the advantages and disadvantages of including multiple markers in the model for QTL mapping. These are the basic principles for selecting appropriate markers in the model as a background control. The possible effects of multiple regression on the overall threshold for an interval test and on the reduction of the degrees of freedom for the test need also to be considered for selecting markers included in the model, as too many markers fitted in the model will reduce significantly the number of degrees of freedom of the test statistic and could reduce significantly the statistical power of detecting QTLs, particularly when the sample size is small.

Bruce Weir, C. Clark Cockerham, Bill Hill, Rebecca Doerge, Dennis Boos, and an anonymous reviewer provided constructive comments on the manuscript. This study was supported in part by grants GM 45344 from the National Institutes of Health and DEB-9220856 from the National Science Foundation.

1. Lander, E. S. & Botstein, D. (1989) *Genetics* **121,** 185–199.
2. Soller, M., Brody, T. & Genizi, A. (1976) *Theor. Appl. Genet.* **47,** 35–39.
3. Haley, C. S. & Knott, S. A. (1992) *Heredity* **69,** 315–324.
4. Martinez, O. & Curnow, R. N. (1992) *Theor. Appl. Genet.* **85,** 480–488.
5. Stuart, A. & Ord, J. K. (1991) *Kendall's Advanced Theory of Statistics* (Oxford Univ. Press, New York), 5th Ed., Vol. 2.